Feature knowledge distillation using group convolutions for efficient plant pest recognition

Kostiantyn Khabarlak[1,*], Ivan Laktionov[1] and Grygorii Diachenko[1]

[1] *Dnipro University of Technology, D. Yavornytskoho Av., 19, Dnipro, 49005, Ukraine*

Abstract

Early plant pest recognition is important to take timely preventive measures to stop pest spread and improve yields. Multiple approaches rely on neural networks to monitor for plant pathology using edge or mobile devices. However, quality of small neural networks is often insufficient. Knowledge distillation can be used to transfer knowledge from large and accurate neural network to a smaller one. In this work we present a novel feature distillation approach based on group convolutions to improve student neural network performance. The final MobileNetV3 network achieves 74.83% classification accuracy on IP102 plant pest dataset. The trained network is fast enough for edge and mobile devices.

Keywords

Pest classification, knowledge distillation, edge computing, mobile neural networks

1. Introduction

Early and accurate plant pest classification facilitates selection of preventive measures to stop pest spread and improves agricultural product yields. Recent research [1], [2] has shown that convolutional neural networks can solve the problem with high efficiency. However, small neural networks that can be inferred on low-power edge devices show lower classification accuracy.

One of effective ways to improve small neural network accuracy is neural network distillation [3]. It is assumed that large neural networks have learned to extract more discriminative features from the dataset than small networks. When the smaller network is tasked not only to discover features, but also to mimic already learnt features from a larger network, the final accuracy is improved. Early plant pathology classification has become more important in agriculture, new approaches for efficient on-device inference have been proposed. Many of the approaches [4], [5] use knowledge distillation to improve the final accuracy.

Distillation can be performed for different parts of the neural network. Output logits [6], weights, attention maps [7] or inner features [8] can be distilled. Good distillation performance can be achieved, for instance, by combining feature and logit distillation. However, when transferring knowledge between networks with different architectures, feature map sizes in the network commonly do not match. Therefore, a special mapping layer is required to adjust layer sizes. Fully connected or convolutional blocks can be used for the mapping layer. It should be noted that the mapping layer might contain a lot of parameters; therefore, learning to map any teacher features to any student features. Thus, the distillation performance is degraded.

In this work we propose a novel feature distillation approach, where group convolutions [9] are used for the mapping layer to distill teacher features to student. The proposed layer consists of 2 group convolutions with a small number of inner convolutional channels. We show that adding feature distillation with the proposed layer to logit distillation, outperforms both logit distillation and feature distillation without group convolution. As a benchmark we use recently proposed IP102 fine-gradient plant pest classification dataset [10], that has large number of pest categories and training images in the field.

ICST-2024: Information Control Systems & Technologies, September, 23 – 25, 2024, Odesa, Ukraine

khabarlak.k.s@nmu.one (K. Khabarlak); laktionov.i.s@nmu.one (I. Laktionov); diachenko.g@nmu.one (G. Diachenko)

0000-0003-4263-0871 (K. Khabarlak); 0000-0001-7857-6382 (I. Laktionov); 0000-0001-9105-1951 (G. Diachenko)

2. Literature Overview

To perform plant pest or plant disease classification, typically, transfer learning is applied. In such a case, an ImageNet-pretrained neural network architecture is taken, which is then finetuned on the downstream task. ImageNet is a large dataset with over 1M images and 1,000 classes to distinguish between. The plant pest or disease datasets are 10−1000 times smaller. The pretrained neural network has learned to extract many common features that can be useful also for plant-related tasks. Thus, during transfer learning the neural network overfits less to the training dataset, resulting in better generalization capabilities.

Larger architectures typically give better results during transfer learning (e.g., large configurations of ResNet [11], ResNeXt [12], EfficientNet V2 [13]). However, they cannot be inferred on edge or mobile devices due high memory and computation power requirements, while sufficiently small neural networks (e.g., MobileNet V2 [14]/V3 [15] family) demonstrate lower accuracy. Thus, finding balance between accuracy and hardware requirements is important research direction plant healthiness-related tasks.

Several approaches are proposed to reduce memory and computation power requirements, like dynamic neural networks [16], [17] and distillation. The latter has originally been proposed in [3], [18]. The distillation approach relies on an assumption that larger neural networks are able to learn more discriminative features, while small neural networks might underfit the data. By enforcing the smaller neural network to mimic output logits or intermediate features of the larger network, network accuracy can be improved.

Multiple distillation-based approaches have been presented for plant healthiness estimation, through early recognition of plant diseases or pests. [19] propose multi-task knowledge distillation approach to improve tomato leaf disease classification accuracy and disease severity estimation. The approach is based on Kullback-Leibler distillation loss joined with attention transfer [7]. In [20] ResNet-50 model is trained using head and feature distillation to improve plant pathology classification accuracy on the Plant Pathology dataset [21]. Authors of [4] propose multistage knowledge distillation method for improving lightweight plant disease detection model. The authors use focal and global distillation for backbone features as proposed in [22], joined with head distillation. The experiments were conducted on the PlantDoc [23] dataset. In [24] the authors investigate the problem of continual learning in agriculture, when the model needs to learn new weed or disease classes incrementally. The common problem in class-incremental learning is that of catastrophic forgetting of the previously learned classes. The authors propose a knowledge distillation-based solution to the problem. Authors of [25] train low-power model for plant disease detection for smart hydroponics using knowledge distillation technique. In [5] a modified neural network architecture is proposed for maize disease detection. Training is performed using channel-wise distillation.

A survey of plant pathology datasets is available in [2]. In this work we use IP102 fine-gradient plant pest classification dataset [10], that has the largest number of pests (102) and annotated in the field images (75,222) among the considered datasets.

Large networks, such as ResNet-50 are often used as knowledge distillation target, while mobile-friendly neural networks, that are suitable for on-device processing are not sufficiently studied in plant pest classification research works. Therefore, in this work we propose a novel group convolution feature mapping layer for mobile-friendly neural networks, that improves the distilled network accuracy of the MobileNetV3 neural network.

3. Materials and Methods

To perform distillation, first teacher network should be selected and trained. In the experiments section we evaluate multiple architectures and select the one with the highest plant pest classification accuracy.

For the target loss used for teacher training the common choice is the cross-entropy loss function:

$$L_T(\hat{y}, y_{true}) = -\frac{1}{N}\sum_{c=1}^{C} y_{true_c} \cdot \log \text{softmax}(\hat{y}_c), \quad (1)$$

where $\hat{y}$ is a vector of convolutional neural network logits, y_{true} is one-hot encoded vector of the true class, N is a number of mini-batch images. The softmax function is defined as follows:

$$softmax(\hat{y}_c) = \frac{\exp \hat{y}_c}{\sum_j \exp(\hat{y}_j)}. \quad (2)$$

In distillation scheme proposed in [3], after the teacher network is trained, the student is trained on a joined targets (Eq. 1) and soft targets (Eq. 3) losses:

$$L_{ST}(\hat{y}_{teach}, \hat{y}_{stud}) = -\frac{T^2}{N}\sum_{c=1}^{C} \text{softmax}\left(\frac{(\hat{y}_{teach})_c}{T}\right) \log\left(\text{softmax}\left(\frac{(\hat{y}_{stud})_c}{T}\right)\right), \quad (3)$$

where $\hat{y}_{stud}$ is a vector of student logits, $\hat{y}_{teach}$ is a vector of teacher logits obtained on the same input image as $\hat{y}_{stud}$, C is the number of classes in the dataset, T is the soft targets temperature.

Multiple works [6], [7] have shown, that by distilling not only the final feature distribution, but also inner features, the final student network performance can be improved. When transferring knowledge between networks with different architectures, feature map sizes in the network commonly do not match. For instance, the last feature map before dense layer of the EfficientNetV2 Large neural network is of size $1280 \times 7 \times 7$, while feature map of MobileNetN3 Large is of size $960 \times 7 \times 7$. Therefore, a special mapping layer is required to adjust layer sizes. Fully connected or convolutional blocks can be used for the mapping layer. However, the mapping layer might contain a lot of parameters; therefore, learning to map any teacher features to any student features. Thus, the distillation performance is degraded.

Group convolutions have been originally introduced in [9]. In the following works [12], [26] the group convolution has been shown as an efficient way of reducing the overall number of parameters and floating-point operations in the neural network architecture with negligeable accuracy loss. In this work we show that group convolution can be efficiently used to perform feature distillation and propose a novel group convolution mapping layer, that is able to improve distillation performance. The proposed layer consists of 2 group convolutions with a small number of convolutional channels between them.

In a convolutional layer the following number of parameters should be trained:

$$N_{parameters} = F_{out}\left(\frac{F_{in} \cdot K^2}{G} + 1\right), \quad (4)$$

where F_{in} is a number of input channels, F_{out} is a number of output channels, K is a kernel size, G is a number of groups in a convolution. $+1$ is given by the bias term. If $G > 1$, the convolution is called group convolution; when $G = 1$ it becomes a conventional convolution. For group convolutions it is required, that F_{in} and F_{out} are divisible by G. As is seen, the number of parameters in distillation mapping layer can be substantially reduced by using group convolutions.

Group convolutions split one large convolution into subgroups; thus, reducing overall number of parameters. Additionally, the number of parameters can be further reduced by using 2 convolutions with a small number of inner channels instead on a single convolution.

To minimize distance between teacher and student networks via the mapping layer, mean squared error loss is used:

$$L_{GML}(\hat{f}_{teach}, \hat{f}_{stud}) = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\left(GML(\hat{f}_{stud})\right)_{ij} - (\hat{f}_{teach})_{ij}\right)^2, \quad (4)$$

where $\hat{f}_{teach}$ is a teacher feature map, $\hat{f}_{stud}$ is a student feature map, H and W are height and width of the output feature map, $GML(\cdot)$ is the proposed group convolution mapping layer.

The final neural network distillation loss with the proposed group convolution mapping layer consists of 3 weighted components (as is shown in the Fig. 1): targets, soft targets, mapping layer losses, and is defined as follows:

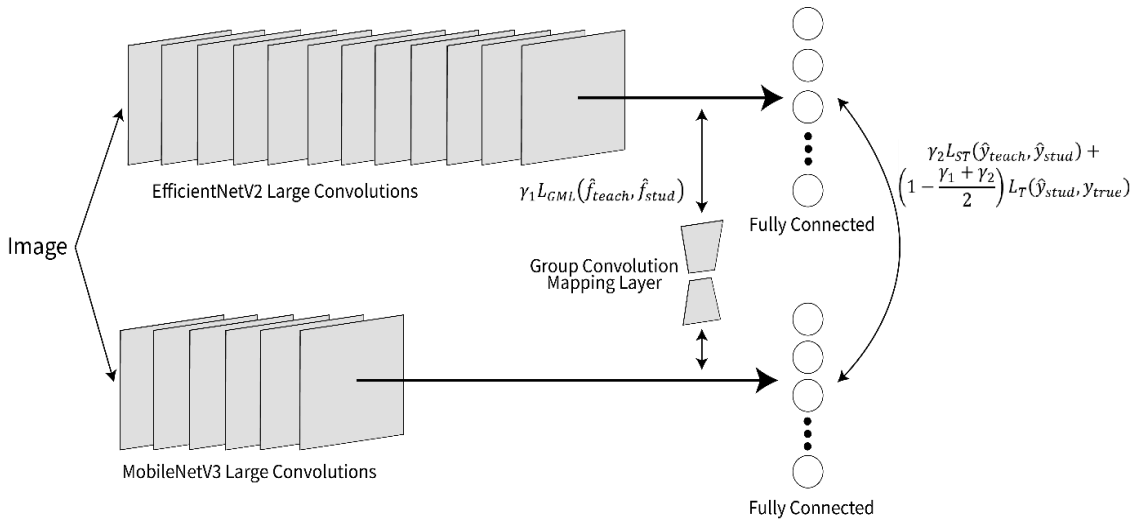


Figure 1: Student network training with group convolution distillation.

$$L(\hat{f}_{teach}, \hat{y}_{teach}, \hat{f}_{stud}, \hat{y}_{stud}, y_{true}) = \gamma_1 L_{GML}(\hat{f}_{teach}, \hat{f}_{stud}) + \gamma_2 L_{ST}(\hat{y}_{teach}, \hat{y}_{stud}) + \left(1 - \frac{\gamma_1 + \gamma_2}{2}\right) L_T(\hat{y}_{stud}, y_{true}), \quad (5)$$

where γ_1, γ_2 are weights of the loss components, algorithm hyperparameters.

4. Experiments

In this work we develop plant pest classification system. The processing should be performed on an edge device with a camera attached with performance level similar or equal to Raspberry PI 4. The monitoring will be performed directly on device. In this work we plan to use MobileNetV3 [15] as target neural network as it was shown as a mobile-friendly neural network with high accuracy [27]. To perform knowledge distillation using the proposed method, first we select teacher network. We have evaluated state-of-the-art neural networks in their largest available configurations: ResNet-152 [11], DenseNet-201 [28], EfficientNet B7 [29], ConvNext Large [30], EfficientNetV2 Large [13]. Transfer learning has been used to train each network from ImageNet weights. Training has been conducted for 20 epochs using Adam gradient descent optimizer with initial learning rate $\alpha = 10^{-3}$ and batch size of 256. Best model weights are selected on validation. Training has been performed on Nvidia RTX 4090 GPU. Results are presented on test set. MobileNetV3 student network has been trained from ImageNet-pretrained weights for 30 epochs with other hyperparameters similar to teacher training. The trained student neural network has been used to perform plant pest classification on Snapdragon 845 mobile CPU. For all experiments images of size 224×224 were used. During training the following augmentations were used to improve model quality: horizontal flip, random resized crop and random rotation. These augmentations were used both for teacher and student networks.

5. Results

First, we train all large convolutional neural networks to select the best network to serve as a teacher. Teacher neural network training results are shown in Table 1. The best result is shown in bold. As can be seen EfficientNetV2 Large has the highest accuracy on the IP102 test set. This network is therefore selected as teacher for all the following experiments. MobileNetV3 training has accuracy of 72.14%, which is lower than that of teacher networks. Hence, it is reasonable to perform feature distillation.

Table 1
Teacher neural network plant disease classification accuracy

Architecture	Test Accuracy (%)	Parameters (millions)
ResNet-152	74.15	60.2
DenseNet-201	73.69	20.0
EfficientNet B7	73.30	66.3
ConvNext Large	75.60	197.8
EfficientNetV2 Large	76.17	118.5

Next, we use grid search to find configuration of the proposed convolutional regressor with the highest accuracy. Searching all hyperparameters in a single stage would require more than 8 days of GPU training. Therefore, the search is performed in 3 stages. The hyperparameter grid search stages are shown in Table 2. Initial sizes of mapping layer convolutions are set to 3×3 kernel, $\gamma_1 = 0.25, \gamma_2 = 0.25, T = 2$. Note, that the number of convolution groups cannot be larger, that number of inner channels. Therefore, for 32 inner channels, 64 groups were not considered. Following stages use several combinations of the best hyperparameters currently found.

Table 2
Hyperparameter search stages

Stage	Parameter Name	Values
I	Inner Channels	32, 64, 128, 256, 512
	Convolution 1 Groups	1, 4, 16, 32, 64
	Convolution 2 Groups	1, 4, 16, 32, 64
II	Convolution 1 Kernel Size	1, 3, 5
	Convolution 2 Kernel Size	1, 3, 5
III	Regressor Weight	0.25, 0.5, 0.75, 1.0
	Distillation Weight	0.25, 0.5, 0.75, 1.0
	Distillation Temperature	2, 5, 7, 10

We have found that the best plant pest classification accuracy is given by the configuration shown in Table 3. Analysis on each hyperparameter importance is presented in the Discussion section.

As can be seen from the Table 4, group convolution distillation improves soft targets distillation. Additionally, using conventional convolutions has worse accuracy than the proposed approach by 0,15 %, while having significantly larger number of parameters allocated for the feature mapping layer: 11,06 versus 0,22 million parameters (excluding the number of parameters in the MobileNetV3 network).

Table 3
Best found hyperparameters

Hyperparameter	Best Value
Inner Channels	64
Convolution 1 Groups	16
Convolution 2 Groups	4
Convolution 1 Kernel Size	3
Convolution 2 Kernel Size	3
Regressor Weight	0.5
Distillation Weight	0.5
Distillation Temperature	10

Table 4
Influence of regressor components on the test accuracy

Configurations	Accuracy (%)	Trainable Parameters (millions)
Teacher (EfficientNetV2 Large)	76.17	117.36
Non-distilled (MobileNetV3 Large)	72.14	4.33
KD (Group Conv Distillation + Targets)	72.44	4.55
KD (Soft Targets + Targets)	74.50	4.33
KD (Conv Distillation + Soft Targets + Targets)	74.68	15.39
KD (Group Conv Distillation + Soft Targets + Targets)	74.83	4.55

6. Discussion

The proposed knowledge distillation approach contains multiple components: training from true labels (targets loss L_T) and distillation (soft targets loss L_{ST}, and L_{GML} loss computed using the proposed group convolution mapping layer). To compute each of the losses and the final training loss L, a number of hyperparameters have to be set. The considered hyperparameters were shown in Table 2. In this section we analyze and discuss influence of these hyperparameters on the final result and the improvement obtained on the IP102 plant pest dataset.

During stage 1 grid search the following hyperparameters were considered: number of channels between the 2 convolutions, number of groups for the first and the second convolutions. Initial value of kernel size has been set to 3 for each of the convolutions. Overall, 116 combinations of hyperparameters were considered. Top 5 configurations based on test set accuracy are shown in Table 5. Also, the best configuration without group convolutions is added.

As can be seen, using 512 or 64 inner channels result in better accuracy. All top configurations use one or both group convolutions (with $G > 1$ in Eq. (4)), instead of the conventional convolution (with $G = 1$). It should be noted that mapping layer of the best configuration in the table adds 4.7 million trainable parameters, which is larger than that of the student network itself (4.33 million parameters). The configuration with conventional convolutions has even larger number of mapping layer parameters (5.1 million) and lower accuracy than other configurations. This large number of parameters might result in the mapping layer learning to perfectly map any teacher features to student features, thus deteriorating feature distillation performance. The second-best configuration with 64 inner channels and both group convolutions has 0.2 million parameters (21.8 times fewer), the third-best has 0.08 million parameters, both with accuracy only slightly worse. Therefore, for the stage 2 hyperparameter search all top-5 configurations were considered.

Table 5
Influence of the number of mapping layer inner channels and the number of groups on accuracy

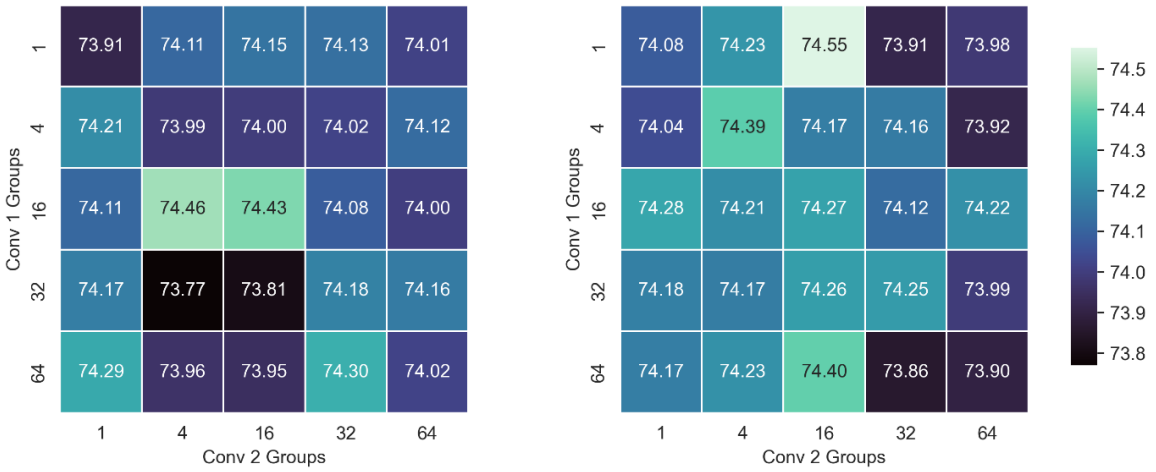
Inner Channels	Conv 1 Groups	Conv 2 Groups	Accuracy (%)	Mapping Layer Params
512	1	16	74.55	4,794,112
64	16	4	74.46	220,224
64	16	16	74.43	81,984
512	64	16	74.40	439,552
512	4	4	74.39	2,582,272
…	…	…	…	…
256	1	1	74.34	5,162,496

In Figure 2 heatmap of the distilled model accuracy versus the number of convolution 1 and 2 groups is shown for the number of inner channels 64 and 512. As can be seen, in both cases using

conventional convolutions (with groups $= 1$) or splitting the convolution into too many groups (e.g. 64 groups for both convolutions) does not give the best results. Choice of 4 or 16 groups seems to be reasonably good. Next, convolution kernel sizes with the highest accuracy for each of the top 5 configurations from Table 4 are searched. The results are shown in Table 6. Clearly, using kernel size of 3 for both convolutions results in the highest accuracy.

Table 6
Influence of regressor convolution kernel sizes on accuracy

Inner Channels	Convolution 1		Convolution 2		Accuracy (%)
	Groups	Kernel Size	Groups	Kernel Size	
512	1	3	16	3	74,55
64	16	3	4	3	74,46
64	16	3	16	3	74,43
64	16	1	4	5	74,40
512	64	3	16	3	74,40



(a) (b)

Figure 2: Heatmap of the distilled model accuracy versus the number of convolution 1 and 2 groups. Inner channels: (a) 64, (b) 512.

Based on the conducted experiments, we estimate group convolution mapping layer hyperparameter importance.

For that, recursive feature elimination with random forest regressor as base algorithm has been used, which is a common approach. As is shown in Figure 2, the number of convolution 1 groups has the highest influence on the test set accuracy, followed by the number of inner channels and convolution 2 groups.

Finally, loss weights γ_1, γ_2 and distillation temperature T with the highest accuracy are searched. In this stage 2 best configurations from previous experiments are considered, namely 512 inner channels with 1 and 16 groups, and 64 inner channels with 16 and 4 groups in the first and second convolutions correspondingly. Overall, 128 combinations of these configurations are searched in stage 3.

The results are shown in Table 3. Therefore, the initial values of $\gamma_1 = 0.25, \gamma_2 = 0.25, T = 2$ are updated to the best $\gamma_1 = 0.5, \gamma_2 = 0.5, T = 10$. Also, configuration with 64 inner channels (and fewer number of trainable parameters) has outperformed the large mapping layer with 512 inner channels.

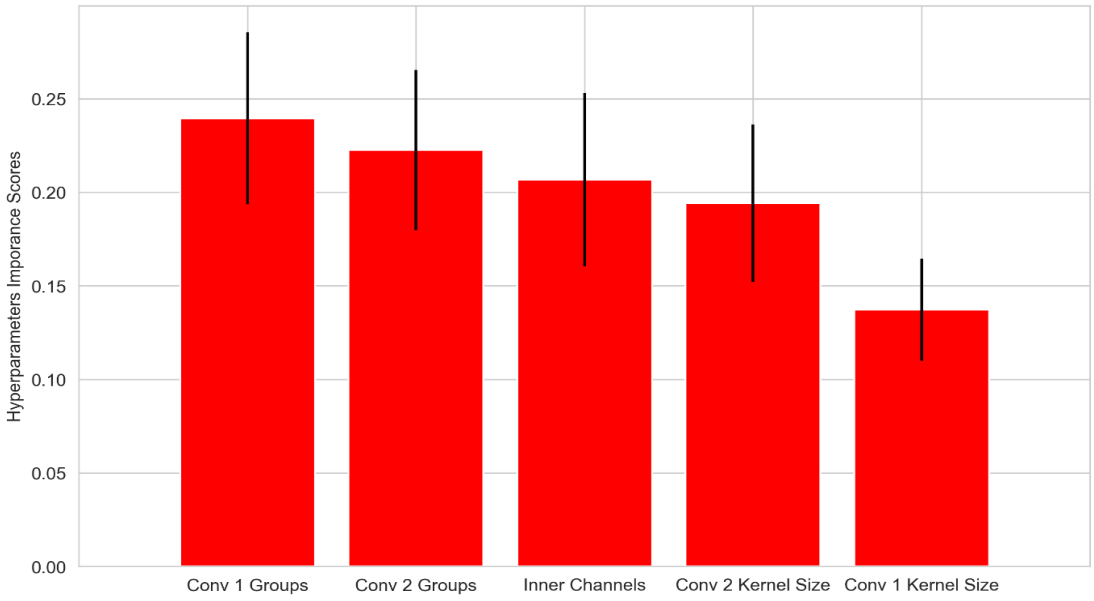


Figure 3: Group convolution mapping layer hyperparameter importance.

Table 7
Influence of weights γ_1, γ_2 and temperature T hyperparameters on accuracy

Mapping Layer				Distillation		
γ_1	Inner Channels	Conv 1 Groups	Conv 2 Groups	γ_2	T	Accuracy (%)
0,50	64	16	4	0,50	10	74,83
0,75	64	16	4	0,50	10	74,76
1,00	512	1	16	0,25	7	74,75
0,50	64	16	4	0,50	7	74,69
0,75	512	1	16	0,50	7	74,65

Finally, we investigate how knowledge distillation influences per class accuracy of plant pest classification. Obviously, it is not possible to show visualization of performance on each of 102 class of the IP102 dataset. Therefore, we sort classes by student test accuracy and visualize every 10th class performance for teacher, MobileNetV3 and the proposed combined knowledge distillation approach of group convolution mapping layer with soft targets as is shown in Figure 4.

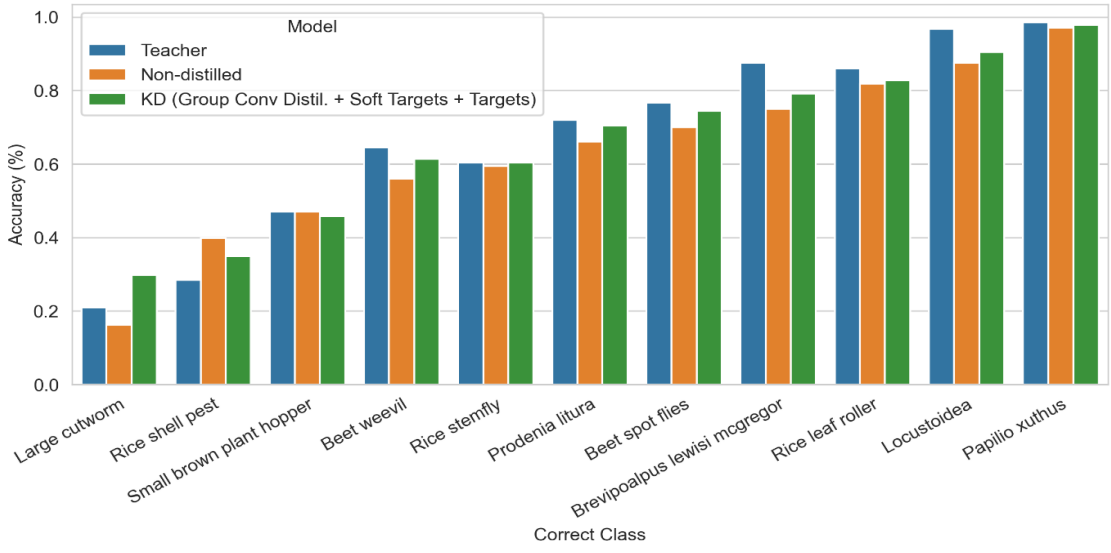


Figure 4: Teacher, non-distilled, distilled student accuracy on some of the IP102 dataset classes.

As can be seen, in most cases distilled model takes intermediate place between teacher and non-distilled MobileNetV3. In rare cases distilled model outperforms (e.g. "Large cutworm") or underperforms (e.g. "Rice shell pest") both teacher and non-distilled models.

7. Conclusions

As has been shown, large number of parameters in the feature mapping layer between teacher and student networks deteriorates performance of knowledge distillation. In this work a novel group-convolution-based feature mapping layer is proposed, that significantly reduces the number of parameters in the mapping layer and improves the student network accuracy. Combined with soft targets distillation, the quality of the MobileNetV3 network is improved from 72.14% to 74.83% in the pest classification task.

Future work will be focused on deploying the trained network on a Raspberry PI 4 stationary greenhouse plant pest monitoring system and improving accuracy of fine-grained plant disease recognition (detection and segmentation) at a large distance.

Acknowledgements

This research was carried out as part of the scientific project "Development of software and hardware of intelligent technologies for sustainable crop production in wartime and post-war" (state registration number 0124U000289) funded by the Ministry of Education and Science of Ukraine at the expense of the state budget.

References

[1] W. B. Demilie, Plant disease detection and classification techniques: a comparative study of the performances, J. Big Data 11 1 (2024) 5. doi: 10.1186/S40537-023-00863-9.

[2] J. Liu and X. Wang, Plant diseases and pests detection based on deep learning: a review, Plant Methods 17 1 (2021). doi: 10.1186/s13007-021-00722-9.

[3] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, CoRR, vol. abs/1503.02531, 2015.

[4] Q. Huang et al., Knowledge distillation facilitates the lightweight and efficient plant diseases detection model, Plant Phenomics 5 (2023). doi: 10.34133/plantphenomics.0062.

[5] Y. Hu, G. Liu, Z. Chen, J. Liu, J. Guo, Lightweight one-stage maize leaf disease detection model with knowledge distillation, Agriculture 13 9 (2023) 1664. doi: 10.3390/agriculture13091664.

[6] Z. Huang, N. Wang, Like what you like: Knowledge distill via neuron selectivity transfer, CoRR, vol. abs/1707.01219, 2017.

[7] S. Zagoruyko and N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, in: ICLR 2017, France, April 24-26, 2017.

[8] J. Kim, S. Park, N. Kwak, Paraphrasing complex network: Network compression via factor transfer, in: NeurIPS 2018, December 3-8, 2018, pp. 2765–2774.

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, December 3-6, 2012, pp. 1106–1114.

[10] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, J. Yang, IP102: A large-scale benchmark dataset for insect pest recognition, in: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE, 2019. doi: 10.1109/cvpr.2019.00899.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, USA, June 27-30, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[12] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, USA, July 21-26, 2017, pp. 5987–5995. doi: 10.1109/CVPR.2017.634.
[13] M. Tan, Q. V. Le, EfficientNetV2: Smaller models and faster training, in: ICML, 18-24 July 2021, PMLR, 2021, pp. 10096–10106.
[14] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: CVPR 2018, USA, June 18-22, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
[15] A. Howard et al., Searching for MobileNetV3, in: ICCV 2019, Korea (South), October 27 - November 2, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
[16] K. Khabarlak, Post-Train Adaptive MobileNet for Fast Anti-Spoofing, in: CEUR Workshop Proceedings, vol. 3156. CEUR-WS.org, 2022, pp. 44–53. URL: http://ceur-ws.org/Vol-3156/keynote5.pdf
[17] K. Khabarlak, Post-train adaptive U-Net for image segmentation, Information Technology: Computer Science, Software Engineering and Cyber Security 2 (2022) 73–78. doi: 10.32782/IT/2022-2-8.
[18] C. Bucila, R. Caruana, A. Niculescu-Mizil, Model compression, in: ACM SIGKDD, USA, August 20-23, 2006, pp. 535–541. doi: 10.1145/1150402.1150464.
[19] B. Liu, S. Wei, F. Zhang, N. Guo, H. Fan, W. Yao, Tomato leaf disease recognition based on multi-task distillation learning, Frontiers in Plant Science 14 (2024). doi: 10.3389/fpls.2023.1330527.
[20] J. Su, S. Anderson, L. Mihaylova, Holistic self-distillation with the squeeze and excitation network for fine-grained plant pathology classification, in: 2023 26th International Conference on Information Fusion, IEEE, Jun. 2023. doi: 10.23919/fusion52260.2023.10224184.
[21] R. Thapa, K. Zhang, N. Snavely, S. Belongie, A. Khan, The Plant Pathology Challenge 2020 data set to classify foliar disease of apples, Applications in Plant Sciences 8 9 (2020). doi: 10.1002/aps3.11390.
[22] Z. Yang et al., Focal and global knowledge distillation for detectors, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, June 18-24, 2022, pp. 4633–4642. doi: 10.1109/CVPR52688.2022.00460.
[23] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, N. Batra, PlantDoc: A dataset for visual plant disease detection," in Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, New York, NY, USA: Association for Computing Machinery, 2020, pp. 249–253. doi: 10.1145/3371158.3371196.
[24] M. P. Fortin, Class-incremental learning of plant and disease detection: Growing branches with knowledge distillation, in: ICCV 2023 - Workshops, France, October 2-6, 2023, pp. 593–603. doi: 10.1109/ICCVW60793.2023.00066.
[25] A. Musa, M. Hassan, M. Hamada, F. Aliyu, Low-power deep learning model for plant disease detection for smart-hydroponics using knowledge distillation techniques, Journal of Low Power Electronics and Applications 12 2 (2022) 24. doi: 10.3390/jlpea12020024.
[26] X. Zhang, X. Zhou, M. Lin, J. Sun, ShuffleNet: An extremely efficient convolutional neural network for mobile devices, in: CVPR 2018, USA, June 18-22, 2018, pp. 6848–6856. doi: 10.1109/CVPR.2018.00716.
[27] K. Khabarlak, L. Koriashkina, Fast facial landmark detection and applications: A survey, Journal of Computer Science and Technology 22 1 (2022) 12–41. doi: 10.24215/16666038.22.E02.
[28] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, USA, July 21-26, 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
[29] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, USA, 9-15 June 2019, pp. 6105–6114.

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, June 18-24, 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.